Chi-Square Testing

Alan B. Gelder
06E:071, The University of Iowa[1]

1 Theoretical Introduction

1.1 From Two to Several Proportions

The Greek letter χ: First, let's introduce the Greek letter χ (called chi and pronounced like "kite" without the 't' sound). The Greek letter χ is not to be confused with the Roman letter x. For instance, Christmas is sometimes referred to as x-mas. However, x-mas is actually derived from χ-mas since χ is the first letter in the Greek spelling of Christ.

A new distribution: So far in this course we have done statistical inference with a single proportion and with two proportions. We will now study a statistical tool for examining a group of proportions. In order to do this we need a new distribution. Recall that we used the normal distribution for one and two proportion problems. Our new distribution, called the χ^2 distribution, is derived from squaring the standard normal distribution. If we have one degree of freedom, then the χ^2 distribution is just the square of the standard normal distribution. If we have two degrees of freedom, then we square two standard normal distributions and add them together to get the χ^2 distribution. In general, for k degrees of freedom, the χ^2 distribution is equal to the sum of k squared standard normal distribution. The χ^2 distribution for two and four degrees of freedom is shown at the top of p. 543 in the text.

Groups of categorical data: Since χ^2 problems deal with groups of proportions, we will be focusing on *categorical* data. (Note that quantitative variables may sometimes be grouped as categorical variables. Age, for example could be grouped into categories as follows: $< 18, 18-25, 26-39, 40-65, > 65$.)

1.2 Defining Hypotheses

Related categorical variables: Fundamentally, in χ^2 problems we are trying to see if two categorical variables are related. We could ask such questions as: Is gender related to the type of car a consumer purchases? Is religion related to occupational

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

choice? Is political affiliation related to what part of the country you grew up in? χ^2 hypothesis testing provides a formal way of addressing such questions.

In words, the null and alternative hypotheses can be stated as follows:
H_0 : *variable 1* and *variable 2* are not related.
H_A : *variable 1* and *variable 2* are related.

In mathematical symbols, the hypotheses are a little more complicated. Consider the following example:

Reese's Pieces example: Suppose that the Hershey's chocolate factory has two machines for packaging Reese's Pieces. There are three colors of Reese's Pieces: yellow, orange, and brown. Workers have been trying to calibrate the two machines so that they will dispense the same proportion of each color into each Reese's Pieces package.

Colors of Reese's Pieces by Packaging Machine

		Yellow	*Orange*	*Brown*
Machine	1	p_{11}	p_{12}	p_{13}
Machine	2	p_{21}	p_{22}	p_{23}

If the workers have correctly calibrated the two machines, then the proportion of yellow Reese's Pieces that machine 1 releases (p_{11}) is equal to the proportion that machine 2 dispenses (p_{21}). The same would be true for orange and brown Reese's Pieces. Hence, the null hypothesis is

$$H_0 : p_{11} = p_{21} \;\&\; p_{12} = p_{22} \;\&\; p_{13} = p_{23}$$

The alternative hypothesis is the logical negation of the null hypothesis. As such, H_A includes all of the possible ways that H_0 could not be true. H_A is given below:

$$H_A : \begin{cases} p_{11} \neq p_{21} \;\&\; p_{12} = p_{22} \;\&\; p_{13} = p_{23}; \text{ or} \\ p_{11} = p_{21} \;\&\; p_{12} \neq p_{22} \;\&\; p_{13} = p_{23}; \text{ or} \\ p_{11} = p_{21} \;\&\; p_{12} = p_{22} \;\&\; p_{13} \neq p_{23}; \text{ or} \\ p_{11} \neq p_{21} \;\&\; p_{12} \neq p_{22} \;\&\; p_{13} = p_{23}; \text{ or} \\ p_{11} \neq p_{21} \;\&\; p_{12} = p_{22} \;\&\; p_{13} \neq p_{23}; \text{ or} \\ p_{11} = p_{21} \;\&\; p_{12} \neq p_{22} \;\&\; p_{13} \neq p_{23}; \text{ or} \\ p_{11} \neq p_{21} \;\&\; p_{12} \neq p_{22} \;\&\; p_{13} \neq p_{23} \end{cases}$$

Intuitively, H_A is saying that somewhere there is at least one pair of proportions that is not equal. Or rather, the two machines are not dispensing the same propor-

tion of at least one color of Reese's Pieces.

In words, the hypotheses are stated as follows:
H_0 : Packaging machines and Reese's Pieces color distribution are *not* related.
H_A : Packaging machines and Reese's Pieces color distribution *are* related.

If H_0 is correct, then there is no way to identify which machine packaged a particular bag of candy based on the colors of the Reese's Pieces inside the bag. This is what is means for the packaging machine to be unrelated to the color distribution.

1.3 Observed and Expected Counts

Like any statistical test, we need to compare our sample distribution (what we observe) with the distribution based on the null hypothesis (what we expect assuming H_0 is true). To illustrate this, let's return to the Reese's Pieces example.

Suppose that each machine is designed to place 60 Reese's Pieces into each package. Workers take a random sample of Reese's Pieces packages from each machine and calculate how many yellow, orange, and brown candies the two machine include in each package, on average. The results are shown below:

Observed Counts

	Yellow	*Orange*	*Brown*	*Total*
Machine 1	12	27	21	60
Machine 2	8	33	19	60
Total	20	60	40	120

Using the row total and the column total, we can calculate how many of each color we would expect to find in a package made by each machine if H_0 is true. Note that 20/120 or 1/6 of all Reese's Pieces are yellow. Since each package from either machine has 60 Reese's Pieces, we would expect to find $(1/6) \times 60 = 10$ yellow candies in each package. Similarly, we would expect to find $(60/120) \times 60 = 30$ orange and $(40/120) \times 60 = 20$ brown Reese's Pieces in each package if H_0 is true.

In general, the formula for the expected count is

$$\text{Expected Count} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

It is helpful to make a table of expected counts when solving a χ^2 problem. The

expected counts for this problem are shown below:

Expected Counts:

	Yellow	*Orange*	*Brown*	*Total*
Machine 1	10	30	20	60
Machine 2	10	30	20	60
Total	20	60	40	120

One way to check your math is that the row and column totals of the expected counts should equal the original row and column totals.

1.4 Chi-Square Contributions and the Chi-Square Statistic

We will now measure how closely our observed counts match our expected counts. If there is not much difference between our sample and what we would expect if H_0 is true, then we will fail to reject H_0. On the other hand, if there is a large difference between the sample and the expected counts, then we reject H_0.

The difference between an observed value and an expected value is calculated using a χ^2 contribution, which is defined as

$$\chi^2 \text{ Contribution} = \frac{(O-E)^2}{E}$$

In this formula, O = Observed Counts, and E = Expected Counts.

Returning to our example, the χ^2 contribution of yellow Reese's Pieces packaged with Machine 1 is $(12-10)^2/10 = 0.40$. Likewise, the χ^2 contribution of orange Reese's Pieces packaged with Machine 1 is $(27-30)^2/30 = 0.30$. The χ^2 contributions for all of the different cells are shown in the following table:

χ^2 Contributions

	Yellow	*Orange*	*Brown*
Machine 1	0.4	0.3	0.05
Machine 2	0.4	0.3	0.05

Note that χ^2 contributions are not calculated for the row totals or column totals.

Finally, the χ^2 statistic is obtained by adding all of the χ^2 contributions together. The formula for the χ^2 statistic is

$$\chi^2 \text{ Statistic} = \sum_{All\ \ Cells} \frac{(O-E)^2}{E}$$

Squaring the difference between the observed counts and expected counts in χ^2 contributions makes it so that observed values below and above their expected values do not cancel each other out in the χ^2 statistic. (The idea of summing squared differences will show up again in regression.)

The χ^2 statistic for the Reese's Pieces problem is

$$\chi^2 = 0.4 + 0.3 + 0.05 + 0.4 + 0.3 + 0.05 = 1.5$$

1.5 Degrees of Freedom

The degrees of freedom in a χ^2 problem can be calculated with the following formula:

$$\text{Degrees of Freedom} = (\#\text{ of Rows} - 1) \times (\#\text{ of Columns} - 1)$$

Note: The number of rows and columns do not include the "Total" row or column.

In the Reese's Pieces example, the degrees of freedom is

$$(2-1) \times (3-1) = 1 \times 2 = 2$$

Intuition for degrees of freedom: Remember that degrees of freedom is a measure of how much information you have to work with. One way to think about degrees of freedom for a χ^2 problems is to imagine that you know all of the totals for your different categories. The degrees of freedom is the minimum number of cells in the χ^2 matrix that you would need to know the values of in order to fill out the values of the remaining cells.

1.6 Using the Chi-Square Table

Unlike the t or the normal distribution, the χ^2 distribution is a one-tailed distribution. Hence, we are only interesting in the area in the right hand side of the distribution (i.e. we always shade to the *right* of our χ^2 statistic). The χ^2 table gives us cut off values where the area in the right hand side of the χ^2 distribution is equal to 0.10, 0.05, 0.01, etc.

Just like using the t-table, the first step to using the χ^2 table is to determine your degrees of freedom. On the first row of the table is a list of tail probabilities. The tail probabilities are the numbers which correspond with your level of tolerance α (e.g. .25, .20, .15, .10, .05, etc.). Where the row for your degrees of freedom and the column for your tail probability intersect is the critical value for your hypothesis test.

If the χ^2 statistic is greater than or equal to the critical value, then we reject H_0. If the χ^2 statistic is less than the critical value, then we fail to reject H_0.

In the Reese's Pieces example, using $\alpha = 0.05$, our critical value is 5.99. However, our χ^2 statistic is only 1.5. Therefore, we fail to reject H_0 since χ^2 stat $= 1.5 < 5.99$. We can interpret this decision as follows: There is not sufficient evidence to show that the packaging machines and the Reese's Pieces color distribution are related.

2 Example: Part 1

Changing Majors

[Adapted from Problem 9.23 in the text]

The College of Science and Mathematics at a university is investigating student retention within the college. For students that transfer out of the college, the college wants to identify if there is any pattern between students' initial major (biology, chemistry, mathematics, or physics) and the major that students transfers to. Data for students that have recently transferred out of the college is given below:

	Engineering	*Management*	*LiberalArts*	*Total*
Biology	13	25	158	398
Chemistry	16	15	19	114
Mathematics	3	11	20	72
Physics	9	5	14	61

Formally test to see if the initial major of students within the College of Science and Mathematics is related to the major that they transfer to. Use the four-steps and $\alpha = 0.05$.

1. Define:
H_A: Initial major of math and science students is *related* to the field that they transfer to.

H_0: Initial major of math and science students is *not related* to the field that they transfer to.

2. Calculate:
Note that in our data engineering, management, and liberal arts do not account for all of the students that transferred out of the math and science majors. We can account for these remaining students with an "other" category. We also do not have column totals. The complete matrix of observed values is written as follows:

Observed Counts

	Engineering	*Management*	*LiberalArts*	*Other*	*Total*
Biology	13	25	158	202	398
Chemistry	16	15	19	64	114
Mathematics	3	11	20	38	72
Physics	9	5	14	33	61
Total	41	56	211	337	645

Based on the observed counts we can find the expected counts. Remember, the expected counts are the numbers that we would expect to find if we only had the row and column totals. **(How do we find expected counts?)**

Expected Counts

	Engineering	*Management*	*LiberalArts*	*Other*	*Total*
Biology	25.30	34.56	130.20	207.95	398
Chemistry	7.25	9.90	37.29	59.56	114
Mathematics	4.58	6.25	23.55	37.62	72
Physics	3.88	5.30	19.96	31.87	61
Total	41	56	211	337	645

The expected counts are what we have if H_0 is true. Now we need to compare that with our actual data. This is done through χ^2 contributions. **Again, the answers are given below, but how do we get them?**

Chi-Square Contributions

	Engineering	*Management*	*LiberalArts*	*Other*
Biology	5.979	2.642	5.937	0.170
Chemistry	10.574	2.630	8.973	0.331
Mathematics	0.543	3.608	0.536	0.004
Physics	6.767	0.017	1.777	0.040

Chi-Square Statistic
$5.979 + 2.642 + 5.937 + 0.170 + 10.574 + 2.630 + 8.973 + 0.331 + 0.543 + 3.608$

$+ 0.536 + 0.004 + 6.767 + 0.017 + 1.777 + 0.040 = 50.527$

Degrees of Freedom
(Rows - 1)×(Columns - 1) = (4-1)×(4-1) = 9

Critical Value
16.92 (Where is this number coming from?)

3. Decide:
Reject H_0 since χ^2 statistic $= 50.527 > 16.92$.

4. Interpret:
There is sufficient evidence that the initial major of students who leave the College of Science and Mathematics is related to the field that they transfer into.

3 Chi-Square as a Gateway Test

The χ^2 test simultaneously checks to see if several sets of proportions are equal. (Now would be a good time to review the mathematical notation for the null and alternative hypotheses in χ^2 tests discussed earlier.) If we fail to reject H_0, then there is no statistical difference between all of the proportions in each set that we are testing. However, if we reject H_0, then all that we are saying is that somewhere there is at least one pair of proportions that are statistically different from each other. We don't know which pair(s) of proportions it is. The chi-square test is therefore a gateway test: it tells us whether or not we need to do some further investigation.

So how do we find out which pairs of proportions are statistically not equal to each other? We use two-proportion hypothesis tests where $H_A : p_i \neq p_j$. Another way to do it is to use two-proportion confidence intervals and see if the confidence interval contains zero (if it contains zero, then there is no difference between the proportions; if it does not contain zero, then the two proportions are different).

Chi-Square Contributions: We could systematically go through and do two-proportion hypothesis tests or confidence intervals on all of the relevant pairs of proportions. However, that may take some time. The χ^2 contributions can be used as a helpful tool for honing in on where differences between proportions may be. The larger the χ^2 contribution of a cell, the more likely that the underlying propor-

tion in that cell will be statistically different from the other proportions that it is being tested against.

Terminology: Suppose that we reject H_0 in the χ^2 test. If there is no statistical difference between a pair of proportions, then they do not support the conclusion of the χ^2 test. If there is a statistical difference between a pair of proportions, then those proportions support the conclusion of the χ^2 test.

4 Underlying Proportions

The underlying proportions in the problem of the students changing majors can be seen below:

	Engineering	*Management*	*LiberalArts*	*Other*
Biology	p_{11}	p_{12}	p_{13}	p_{14}
Chemistry	p_{21}	p_{22}	p_{23}	p_{24}
Mathematics	p_{31}	p_{32}	p_{33}	p_{34}
Physics	p_{41}	p_{42}	p_{43}	p_{44}

There are two ways to view these underlying proportions, and each way is based on a corresponding set of populations. First, we have the populations of students that transfer from the College of Science and Mathematics to a particular major.

Population 1: All students that transfer to Engineering from the college
Population 2: All students that transfer to Management from the college
Population 3: All students that transfer to Liberal Arts from the college
Population 4: All students that transfer to some other major from the college

With these populations, the sample proportions in the the χ^2 problem are the following:

	Engineering	*Management*	*LiberalArts*	*Other*
Biology	13/41	25/56	158/211	202/337
Chemistry	16/41	15/56	19/211	64/337
Mathematics	3/41	11/56	20/211	38/337
Physics	9/41	5/56	14/211	33/337

Using this set-up, the null hypothesis expresses that the proportion of all students that transfer to Major A who started in Major C is equal to the proportion of all students that transfer to Major B who also started in Major C (e.g. the proportion of students transferring to engineering from biology is the same as the proportion

of students transferring to management from biology). Mathematically, the null hypothesis can be written as follows:

$$H_0 : \begin{cases} p_{11} = p_{12} = p_{13} = p_{14} \text{ and} \\ p_{21} = p_{22} = p_{23} = p_{24} \text{ and} \\ p_{31} = p_{32} = p_{33} = p_{34} \text{ and} \\ p_{41} = p_{42} = p_{43} = p_{44} \end{cases}$$

The second way to define the populations is in terms of the students that transfer out of each major.

Population A: All Biology students that transfer out of the college
Population B: All Chemistry students that transfer out of the college
Population C: All Mathematics students that transfer out of the college
Population D: All Physics students that transfer out of the college

Using this set of populations, the underlying proportions are as follows:

	Engineering	*Management*	*LiberalArts*	*Other*
Biology	13/398	25/398	158/398	202/398
Chemistry	16/114	15/114	19/114	64/114
Mathematics	3/72	11/72	20/72	38/72
Physics	9/61	5/61	14/61	33/61

The null hypothesis here implies that of all the students that transfer to Major A, the proportion of transfer students from Major B is equal to the proportion of transfer students from Major C. For instance, of all the students that transferred to engineering, the proportion of transfer students that came from biology is equal to the proportion of transfer students that came from physics. The formal mathematical statement of the null hypothesis is given below:

$$H_0 : \begin{cases} p_{11} = p_{21} = p_{31} = p_{41} \text{ and} \\ p_{12} = p_{22} = p_{32} = p_{42} \text{ and} \\ p_{13} = p_{23} = p_{33} = p_{43} \text{ and} \\ p_{14} = p_{24} = p_{34} = p_{44} \end{cases}$$

5 Example: Part 2

Taking a closer look at pairs of proportions

1. Find and interpret a 95% confidence interval for the difference in the proportions of biology and chemistry students who transferred to engineering.

2. Find and interpret a 95% confidence interval for the difference in the proportions of mathematics and physics students who transferred to liberal arts.

3. Find and interpret a 95% confidence interval for the difference in the proportion of students transferring to engineering from physics and the proportion of students transferring to management from physics?

4. Do these confidence intervals support the conclusion of the χ^2 test? Why or why not?

Notes on the answers

The first question compares the members of Population A who transfer to engineering with the members of Population B who transfer to engineering. So $\hat{p}_{11} = 13/398$ and $\hat{p}_{21} = 16/114$. The 95% confidence interval for $p_{11} - p_{21} = (-0.1738, -0.0416)$. Hence, we are 95% confident that between 4.2% and 17.4% fewer students transfer from biology to engineering than from chemistry to engineering. Since the confidence interval does not contain zero (i.e. since there is a statistical difference between p_{11} and p_{21}) then the confidence interval supports the desicion of the χ^2 test. (Remember, there should be at least one difference in a pair of proportions since rejected H_0.)

The third question compares members of Population 1 who started as physics majors with members of Population 2 who also started as physics majors. The sample proportions are $\hat{p}_{41} = 9/41$ and $\hat{p}_{42} = 5/56$, and a 95% confidence interval for the difference in proportions $p_{41} - p_{42}$ is $(-0.0168, 0.2773)$. So, we are 95% confident that the proportion of students that transfer to engineering from physics is between 1.7% lower and 27.7% higher than the proportion of students that transfer to management from physics. Zero is in the confidence interval here, which indicates that there is no statistical difference between p_{41} and p_{42}. Therefore, this comparison does not support the result of the χ^2 test.